Machine Learning
Project

Comparative Analysis of
Different Machine Learning Models
For Activity Recognition
Using
MHEALTH Dataset

Name : Tasneem Mohammed
ID : 202201798

11/5/2024

1. Introduction

This project aims to compare the performance of different machine learning models for activity recognition using the MHEALTH dataset. The models evaluated include K-Nearest Neighbors (KNN), Support Vector Machine, Neural Networks and Logistic Regression.

2. Data Preprocessing & Visualization:

Since the data is not balanced , where some activity classes had fewer data points compared to others.

So I made the following steps :

- I addressed the imbalanced classes by creating a balanced sample.
- Shuffled the sample to be representative.
- Dropped unnecessary columns like the subject column.
- Standardized features using StandardScaler.
- Split data into training and testing sets.

3. Model Training

- **KNN :** Hyperparameter tuning was performed between the values 3, 5 and 7 to find the optimal number of neighbors (k) .
- **SVM:** implemented SVM with RBF kernel, suitable for non-linear data.
- **NN:** a sequential NN architecture with (2) hidden layers was employed. The hidden layers with RelU activation function, and softmax in the output function.
- **Logistic Regression:** GridSearchCV was employed to find the best regularization parameter (C) for the logistic regression model.

4. Model Evaluation

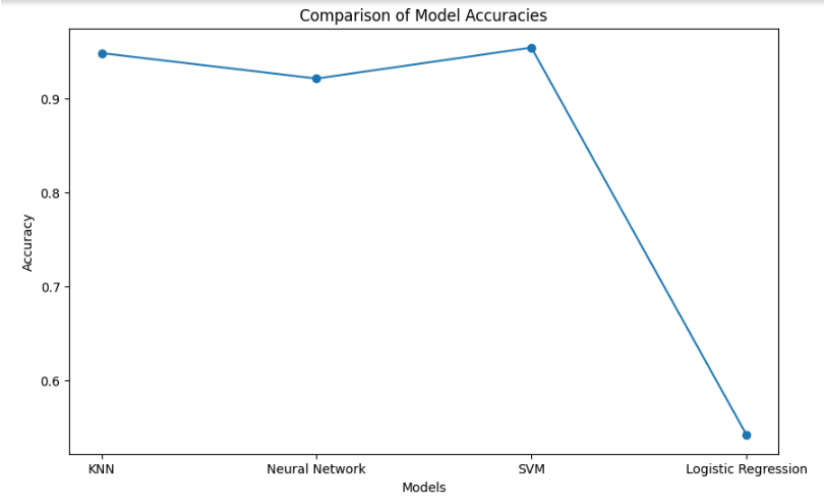
Accuracy, confusion matrix, and classification report were used for evaluation.

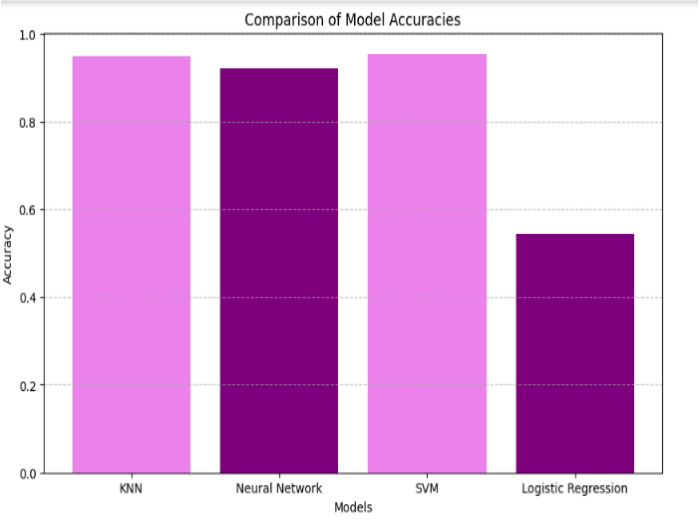
5. Results Analysis

a. Interpretation of Model Predictions

The accuracies achieved by the models indicate their ability to correctly classify activities in the mobile health dataset. The results:

- **KNN** : Achieved an accuracy of 0.9482, which can accurately predict activities but based on their similarity to past examples in the training data.
- **SVM** : Achieved the **highest** accuracy (0.9540), showing its effectiveness in learning decision boundaries to separate different activity classes and find the perfect hyperplane.
- **Neural Network:** Achieved an accuracy of 0.9211, indicating its capability to learn complex patterns.
- **Logistic Regression**: Achieved the lower accuracy (0.5426), this model is not well-suited for this multi-class classification task.





b. Identification of the Best-Performing Model

Based on the evaluation metric (accuracy), the **SVM** model emerges as the **best performer** with an accuracy of **95.4%.**

c. Discussion on Strengths and Weaknesses of Each Model

- KNN: Easy to use, interpretable, but sensitive to parameter k and can be slow for large datasets.
- SVM: Handles complex data well, but can be computationally expensive and lacks interpretability.
- Neural Network: Powerful for complex patterns, but can lead to overfitting and requires careful hyperparameter tuning (black box model).
- Logistic Regression: Fast and interpretable, but limited to linear relationships and may not be ideal for this task (multi classification).

d. Insights into Performance Variation

The performance variation across the models can be attributed to several factors:

- Model complexity: more complex models (SVM, Neural Network) can learn intricate patterns (potentially higher accuracy) but there is a risk of overfitting.
- Data suitability: Logistic Regression might struggle with this multi-class .data
- Hyperparameter tuning: SVM and Neural Network performance are sensitive to hyperparameter choices.

6. Conclusion

SVM emerged as the best model (95.4% accuracy) for activity recognition using the MHEALTH dataset.

Its ability to handle complex data exceeded KNN and Neural Networks (good accuracy), while Logistic Regression struggled with the multi-class classification task.